Invisible Ink, Visible Lies: How Production Watermarking Causes LLMs to Hallucinate

Haocheng Ye[1*] **Aoting Hu**[2*] **Xinwei Zhang**[3†]
Xunzhu Tang[4] **Shuchao Pang**[1†] **Minhui Xue**[5]

[1]Nanjing University of Science and Technology
[2]Anhui University of Technology
[3]The Hong Kong Polytechnic University
[4]University of Luxembourg
[5]CSIRO

Abstract

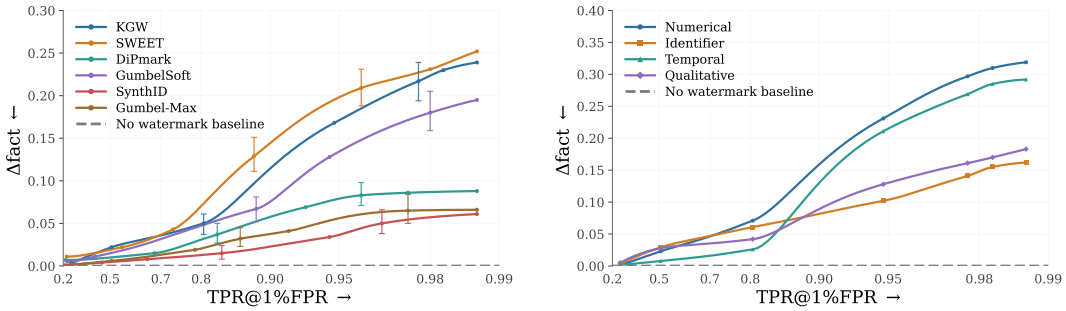
Text watermarking helps identify AI-generated content, but its effect on factual reliability remains underexplored. In this paper, we study *watermarking hallucination*: factual errors induced or amplified by watermarking even when the required evidence is present in the context and the unwatermarked model can answer correctly. Using a controlled retrieval-augmented generation setting, we compare unwatermarked and watermarked generations under the same context, query, and decoding configuration, and quantify their factual accuracy decrease. Across six representative watermarking methods, including KGW, SWEET, DiPmark, GumbelSoft, Gumbel-Max, and SynthID watermarking, we consistently observe watermark-induced hallucination. Watermarked outputs can remain fluent while introducing factual errors. We attribute this failure mode to two mechanisms: ① token perturbations in the current-step arising from method-specific reweighting or keyed sampling, and ② prefix-induced attention drift, which accumulates through autoregressive decoding and weakens later attention to the factual context. Motivated by this analysis, we propose two plug-in interventions at the token and attention levels that can be integrated into existing watermarking methods to improve factuality. At a matched TPR of 0.90 at 1% FPR, combining the two interventions reduces factual errors by approximately 90% relative to watermark-only decoding while preserving fluency and comparable decoding efficiency. Overall, this work highlights factuality as a first-class criterion in watermark evaluation, alongside detectability and robustness, and calls for careful factuality validation before deploying watermarks in fact-critical applications. Code is available at `https://github.com/Haocheng-NJUST/WM-Hallucination`.

1 Introduction

Large language models (LLMs) increasingly generate long-form text in settings where provenance and factual reliability both matter, including clinical notes, legal summaries, financial reports, and public-facing assistants [1–4]. Text watermarking offers a way to identify model-generated content by embedding a statistical signal into the decoding process [5–7]. Some watermarking methods explicitly reweight token probabilities, while others preserve the model distribution under prescribed randomization. In both cases, the realized watermark-conditioned token choices can alter generation trajectories and factual outcomes.

[*]Equal contribution.
[†]Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Factual accuracy decrease across methods.

(b) Fact-type breakdown for KGW [5].

Figure 1: **Factual accuracy decrease under watermarking.** (a) Net factual accuracy decrease, Δ_{fact}, as a function of watermark detection true-positive rate (TPR) at a fixed false-positive rate (FPR) of 1%, across six watermarking methods. Error bars indicate paired bootstrap confidence intervals. (b) Breakdown of Δ_{fact} by fact type for KGW.

Existing watermark evaluations typically assess watermarked text from two perspectives: *watermark effectiveness*, measured by detectability and robustness, and *output quality*, measured by perplexity, semantic similarity, and readability [8–10]. These criteria are necessary for deployment, but they often ignore whether the watermarked output preserves facts supported by the prompt or retrieved context. Recently, Hastuti et al. [11] showed that watermarked medical text can remain fluent while changing a dosage, date, identifier, or legal citation. They show that factuality can degrade even when coherence-oriented quality measures remain acceptable. However, their study does not isolate the factual degradation caused by watermarking itself, explain why such errors arise during decoding, or propose a mitigation strategy.

We refer to this failure mode as *watermarking hallucination* and study it in a controlled retrieval-augmented generation (RAG) setting [12, 13]. Although retrieval and grounding can reduce LLM hallucinations [14–16], watermarking can introduce factual errors even when the relevant facts are available in the supplied context and the same model answers correctly without watermarking. We quantify this effect by the net factual accuracy decrease, $\Delta_{\text{fact}} = \text{Acc}_{\text{clean}} - \text{Acc}_{\text{wm}}$, where $\text{Acc}_{\text{clean}}$ and Acc_{wm} denote factual accuracy for paired unwatermarked and watermarked outputs, respectively, evaluated under matched conditions. Positive values indicate lower factual accuracy under watermarking. Our evaluation covers six watermarking methods: KGW [5], SWEET [17], DiPmark [18], GumbelSoft [19], SynthID-style watermarking [7], and Gumbel-Max [6]. As shown in Figure 1(a), factual accuracy decreases across the evaluated methods, with larger decreases generally observed at higher watermark detection rates. The KGW breakdown in Figure 1(b) further shows that numerical and temporal facts exhibit larger decreases than the other fact types considered.

To investigate the cause of this degradation, we compare factual accuracy between watermarked and non-watermarked generation, and analyze output logits and internal attention patterns. We identify two coupled pathways. At the token level, distortionary methods can directly bias the decoder against a fact-consistent token. At the prefix level, earlier watermark-conditioned token choices can alter the generated prefix and generation trajectory, thereby reducing subsequent attention to fact-bearing context. This *prefix-induced drift* is more difficult to address and has broader consequences than a local token substitution, because even non-factual prefix changes can accumulate and increase the likelihood of later factual errors.

Motivated by these findings, we introduce two fact-preserving interventions, which can be integrated as plug-in components into existing watermarking methods to improve the factual accuracy of watermarked text. Specifically, Fact-Preserving Token Intervention (FPTI) limits watermark-conditioned changes at factual positions, whereas Fact-Preserving Attention Intervention (FPAI) targets prefix-mediated attention loss by restoring attention toward fact-bearing context. At TPR@1%FPR $= 0.90$, their joint application reduces Δ_{fact} by around 90% relative to watermark-only decoding, substantially improving the trustworthiness of watermarked generation. Nevertheless, a residual gap remains under strong watermarking. This gap reflects a hard constraint: even when a method preserves the model distribution under its prescribed randomization, a realized watermark-conditioned trajectory can still diverge from the narrow path required by exact factual grounding.

Overall, the paper makes three contributions.

- We identify *watermarking hallucination* as a factual reliability failure mode distinct from surface-quality degradation and ordinary hallucination, and quantify watermark-induced factual loss under paired controlled RAG evaluation.
- We attribute watermarking hallucination to two coupled decoding-level pathways: token-level perturbations, arising through systematic bias in distortionary methods, and prefix-induced attention interference, which weakens later reliance on factual context.
- We propose two complementary interventions that protect factual token selection and restore attention to supporting evidence, reducing factual degradation under watermarking across multiple watermarking methods and model families.

2 Problem Formulation

2.1 Context-Grounded Generation

We study factual generation under a context-grounded setting. Let x denote the user prompt together with the retrieved or provided factual context. The model generates an answer $y_{1:T}$ autoregressively:

$$p_\theta(y_{1:T} \mid x) = \prod_{t=1}^{T} p_\theta(y_t \mid x, y_{<t}). \tag{1}$$

At each decoding step t, the model assigns a token sampling score $z_t(y_t \mid x, y_{<t})$ to each candidate token y_t. In a factual task, some answer spans are expected to be supported by the context. We call these spans *factual decision points*. At such a point, the model must select a token that is consistent with the evidence in x, such as an entity name, date, number, identifier, or relation.

In standard generation, factual errors may arise when the next-token distribution is not sufficiently anchored to the relevant evidence. This can happen because of missing parametric knowledge, conflict between model memory and the provided context, or gradual drift away from the evidence during generation [14–16]. In RAG-based generation, the supplied context is expected to reduce this uncertainty. Therefore, factuality depends not only on whether the final string looks fluent, but also on whether the relevant evidence remains active when each factual span is generated.

2.2 Decoding-Time Watermarking

We focus on decoding-time watermarking, which modifies token selection during generation without changing the model parameters [5, 17–19]. Let $\tilde{y}_{1:T}$ denote the sequence generated under watermarking. At step t, the model computes the token sampling scores $z_t(y_t \mid x, \tilde{y}_{<t})$ over candidate tokens y_t, conditioned on the watermarked prefix $\tilde{y}_{<t}$. The watermarking operator $\mathcal{W}_t$ transforms these into watermarked scores:

$$\tilde{z}_t = \mathcal{W}_t(z_t(\cdot \mid x, \tilde{y}_{<t}), \tilde{y}_{<t})\,. \tag{2}$$

Here, $\cdot$ denotes scores over the full vocabulary; dependence on the watermark key and sampling randomness is omitted for brevity. We define the resulting token-level score perturbation as

$$b_t(y_t; \tilde{y}_{<t}) = \tilde{z}_t(y_t) - z_t(y_t \mid x, \tilde{y}_{<t}). \tag{3}$$

The next token is selected from the watermarked scores using the method-specific sampling rule. These scores depend on a prefix already shaped by earlier watermarking steps, rather than on the prefix of an unwatermarked generation. Watermarking therefore affects the next token both by changing its current sampling score and by changing the preceding tokens on which that score depends.

2.3 Watermarking Hallucination

We define *watermarking hallucination* as factual degradation induced or amplified by the watermarking process. Let $F(\cdot, x)$ be a factuality evaluator that checks whether an answer is supported by the context x, where $F(y, x) = 1$ denotes a factual answer and $F(y, x) = 0$ denotes a factual error. For a paired generation under the same prompt and decoding configuration, let y^0 be the

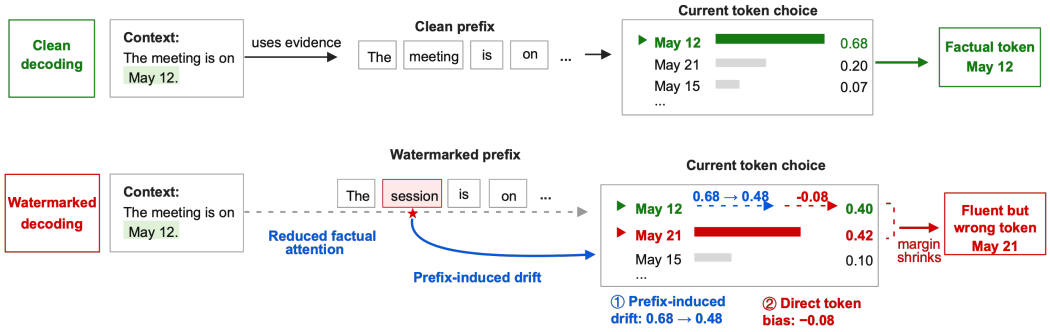


Figure 2: **Decomposition of watermarking hallucination.** Clean decoding preserves the factual margin by following the fact-bearing context. Watermarked decoding can first induce prefix drift, e.g., changing "meeting" to the fluent alternative "session", which weakens factual reliance and lowers the score of the factual token. The current-step score perturbation then further compresses the margin, allowing a fluent distractor such as *May 21* to overtake the evidence-supported token *May 12*.

unwatermarked output and $\tilde{y}$ be the watermarked output. A watermarking hallucination occurs when the base generation is factual but the corresponding watermarked generation is not:

$$F(y^0, x) = 1, \qquad F(\tilde{y}, x) = 0. \tag{4}$$

This paired definition isolates factual errors introduced when watermarking shifts an otherwise context-supported generation toward a factually inconsistent alternative. Since watermarking may also turn an incorrect base generation into a factual one, we quantify its aggregate effect using the *factual accuracy decrease*:

$$\begin{aligned}\Delta_{\text{fact}} = \frac{1}{|\mathcal{D}|} \sum_{(x,q)\in\mathcal{D}} \Big(&\mathbb{I}\big[F(y^0, x) = 1 \wedge F(\tilde{y}, x) = 0\big] \\ &- \mathbb{I}\big[F(y^0, x) = 0 \wedge F(\tilde{y}, x) = 1\big] \Big).\end{aligned} \tag{5}$$

The first term counts answers that change from correct to incorrect under watermarking, while the second counts those that change from incorrect to correct. Their difference, Δ_{fact}, measures the net factual accuracy decrease relative to unwatermarked generation. Positive values indicate degradation, and negative values indicate improvement.

3 Mechanism Analysis

3.1 A Decomposition of Watermarking Hallucination

Watermarking hallucination can be understood as a recursive compression of the factual margin. At a factual decision point, let $y_t^\star$ denote the fact-consistent token supported by the context, and let V_t be the effective candidate set under the sampling rule. Under the clean generation trajectory $y^0_{<t}$, we define the factual robustness margin as

$$\epsilon_t(y_t^\star) = z_t(y_t^\star \mid x, y^0_{<t}) - \max_{y_t \in V_t \setminus \{y_t^\star\}} z_t(y_t \mid x, y^0_{<t}). \tag{6}$$

This margin measures how far the evidence-supported token leads its strongest plausible alternative before watermarking.

As illustrated in Figure 2, watermarking can erode this margin through two coupled effects. First, earlier watermark-conditioned token choices become part of the prefix and shift the model's later distribution. Even if such choices are fluent and semantically close to the clean prefix, they can weaken reliance on the original fact-bearing context. Second, the current watermarking rule can alter the effective selection scores of candidate tokens or the realized token selection, favoring a fluent but factually incorrect competitor over the evidence-supported token. Distortionary methods systematically reweight token probabilities. Distribution-preserving methods preserve the marginal

distribution but can still change individual token choices. Thus, factual errors can arise when prefix-induced drift and the current-step token perturbation jointly reverse the local margin.

Formally, let $y^0_{1:T}$ and $\tilde{y}_{1:T}$ denote the clean and watermarked generation trajectories, and let $\tilde{z}_t(y_t)$ denote the watermarked score defined above. The comparison below requires a common score representation and scale within each method. At step t, its change relative to clean decoding can be decomposed as

$$\tilde{z}_t(y_t) - z_t(y_t \mid x, y^0_{<t}) = \underbrace{b_t(y_t; \tilde{y}_{<t})}_{\text{① token-level perturbation}} + \underbrace{z_t(y_t \mid x, \tilde{y}_{<t}) - z_t(y_t \mid x, y^0_{<t})}_{\text{② prefix-induced drift}}. \tag{7}$$

The first term is the current-step score perturbation defined in Equation (3); it does not necessarily imply a systematic bias in the marginal token distribution. The second term captures the score change caused by conditioning on the watermarked prefix rather than the clean prefix. This prefix-induced drift is recursive: a watermark-conditioned token changes the prefix, which shifts the sampling scores at subsequent steps.

This view explains why watermarking can introduce factual errors without causing visible incoherence. The incorrect token may be locally fluent and semantically close to the correct one, but factual grounding often requires selecting a narrow set of exact entities, numbers, dates, and identifiers. Watermarking hallucination arises when the current-step token perturbation, prefix-induced drift, or their combination erases the margin in Equation (6).

3.2 Empirical Evidence of the Two Channels

The decomposition in Equation (7) suggests two channels through which watermarking affects factual generation.

① **Direct token-level perturbation.** At the current decoding step, watermarking can affect token selection through method-specific reweighting or keyed sampling rules. Because the watermark signal is not explicitly fact-aware, a fact-consistent token may receive weaker support or lose the realized selection to a plausible but incorrect competitor. This channel explains local substitutions in names, dates, numbers, and identifiers while preserving fluency.

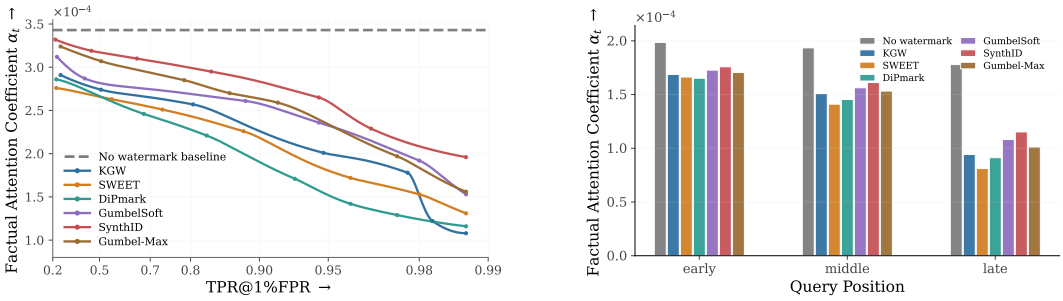
② **Prefix-induced attention interference.** Prefix-induced drift captures the effect of conditioning on the watermarked prefix rather than the clean prefix. In context-grounded generation, we measure one observable form of this drift by tracking how much the current decoding step attends to the fact-bearing tokens in the retrieved context. Let $\mathcal{F}(x) \subseteq \{1, \ldots, |x|\}$ denote the positions of input-side context tokens that support the target fact, and let $A_t(j)$ be the attention weight from the current decoding position t to context token j. For example, in Figure 2, $\mathcal{F}(x)$ contains the context tokens corresponding to "May 12", because they provide the evidence for the next factual token. We define the factual attention coefficient as

$$\alpha_t = \sum_{j \in \mathcal{F}(x)} A_t(j). \tag{8}$$

Thus, α_t measures attention to the supporting evidence in the RAG context. A watermarked prefix such as "The session is on" can change the current decoding state and reduce attention to the context evidence "May 12", making the model more likely to follow parametric priors or fluent distractors.

Together, these two channels explain why watermarking hallucination can occur without visible incoherence: watermarking searches within the broad space of fluent continuations, whereas factual grounding often requires a narrow path through exact entities, numbers, dates, and relations.

Empirical evidence. To isolate prefix-induced interference from the current-step token perturbation, we watermark ordinary answer tokens but disable the watermark perturbation when the next token is factual, i.e., $b_t(y^\star_t; \tilde{y}_{<t}) = 0$. We then measure α_t over the supporting context tokens. As shown in Figure 3(a), α_t consistently decreases as the detectability of the already generated prefix increases. Since the factual token itself is not directly watermarked, this trend supports a prefix-mediated attention effect.



(a) Attention to facts during watermarking

(b) Watermarking hallucination avalanche

Figure 3: **Attention-level interference under watermarking.** (a) Factual attention coefficient α_t decreases as watermark detectability increases, measured by TPR@1%FPR, across six representative methods. (b) At a fixed detectability level, $\mathrm{TPR@1\%FPR} = 0.9$, factual attention is lower when the same target query is answered later, indicating that attention degradation accumulates over the watermarked prefix.

We further test whether this effect accumulates over longer generations by fixing the evidence block and target query while varying only the target query position:

$$\begin{aligned} P_{\text{early}} &= [\mathcal{C};\, Q^\star, Q_1, Q_2], \\ P_{\text{middle}} &= [\mathcal{C};\, Q_1, Q^\star, Q_2], \\ P_{\text{late}} &= [\mathcal{C};\, Q_1, Q_2, Q^\star]. \end{aligned} \tag{9}$$

All other factors, including the supporting evidence $E^\star$, target query $Q^\star$, prompt template, sampling parameters, and matched detectability level, are fixed. Figure 3(b) shows that factual attention is lower when the same query is answered later under watermarking, while the unwatermarked baseline shows a much smaller position effect. This pattern indicates a *watermark avalanche*: early watermark-conditioned choices alter the prefix, which weakens later attention to evidence and increases the risk of factual errors. Implementation details for attention measurement and detectability alignment are provided in Appendix B.4.

4 Fact-Preserving Interventions

Mechanism analysis identifies two sources of watermarking hallucination: current-step token perturbations induced by the watermarking rule and prefix-induced drift accumulated through autoregressive decoding. We therefore introduce two targeted interventions. Fact-Preserving Token Intervention (FPTI) neutralizes watermark perturbations on fact-bearing output tokens, while Fact-Preserving Attention Intervention (FPAI) compensates for the reduced attention to fact-bearing context. Both interventions are designed as plug-in components for decoding-time watermarking methods: FPTI operates on output logits or sampling scores, whereas FPAI requires access to attention logits during inference.

4.1 Fact-Preserving Token Intervention

FPTI targets the token-level perturbation term in Equation (7). Given the input context x, we construct a vocabulary-level factual token set $\mathcal{V}_{\text{fact}} \subseteq \mathcal{V}$, which contains tokens corresponding to entities, dates, numerical values, identifiers, and other fact-bearing spans in the context. In our implementation, these tokens are extracted automatically from the RAG context [20] using named-entity recognition [21] and deterministic pattern-based matching [22].

During decoding, FPTI removes the current-step watermark perturbation from factual tokens while leaving the watermarking rule unchanged elsewhere:

$$\tilde{z}_t^{\text{FPTI}}(y_t) = \begin{cases} z_t(y_t \mid x, \tilde{y}_{<t}), & y_t \in \mathcal{V}_{\text{fact}}, \\ z_t(y_t \mid x, \tilde{y}_{<t}) + b_t(y_t; \tilde{y}_{<t}), & y_t \notin \mathcal{V}_{\text{fact}}. \end{cases} \tag{10}$$

Here $b_t(y_t; \tilde{y}_{<t})$ is the score perturbation defined in Equation (3). FPTI prevents this rule from directly changing the effective selection scores of tokens that carry factual content. Therefore, improvements

from FPTI estimate how much factual degradation is attributable to current-step token perturbations, while any remaining gap indicates the role of prefix-induced drift.

4.2 Fact-Preserving Attention Intervention

FPAI targets an attention-mediated component of prefix-induced drift. Unlike FPTI, it does not modify the output-level watermark perturbation $b_t(y_t; \tilde{y}_{<t})$. Instead, it restores the model's access to fact-bearing evidence when the watermarked prefix reduces factual attention.

Let $\mathcal{F}(x) \subseteq \{1, \ldots, |x|\}$ denote the positions of fact-bearing tokens in the input context. At decoding step t, let $s_t(j)$ be the pre-softmax attention logit assigned to context token j. FPAI adds a compensation term λ_t to fact-bearing context positions:

$$s_t^{\mathrm{FPAI}}(j) = \begin{cases} s_t(j) + \lambda_t, & j \in \mathcal{F}(x), \\ s_t(j), & j \notin \mathcal{F}(x), \end{cases} \qquad A_t^{\mathrm{FPAI}} = \mathrm{softmax}\big(s_t^{\mathrm{FPAI}}\big)\,. \tag{11}$$

The compensation strength λ_t is calibrated to offset the expected attention drop observed under watermarking. Specifically, we model the decline of the factual attention coefficient α_t as a function of prefix detectability and generation position:

$$\lambda_t = \phi(Z_{<t},\, r_t)\,, \tag{12}$$

where $Z_{<t}$ denotes the watermark score of the generated prefix and r_t denotes the relative position of the current factual token. The calibration objective is to restore factual attention toward its clean-decoding level, i.e., $\alpha_t^{\mathrm{FPAI}} \approx \alpha_t^{\mathrm{clean}}$, without changing the output-level watermark rule. In practice, $\phi(\cdot)$ is fitted on held-out calibration data.

Combining two interventions. The two interventions are complementary and can be applied jointly. FPAI restores the evidence-access path by increasing attention to fact-bearing context tokens, while FPTI prevents the subsequent watermark rule from directly perturbing fact-bearing output tokens. Thus, the combined intervention first improves the contextual basis for factual prediction and then protects the factual token choice itself. This allows the watermark to remain active on non-factual tokens while reducing both prefix-induced drift and direct token-level perturbations.

5 Experiments

5.1 Experimental Setup

Data and task. We evaluate watermarking hallucination in a controlled RAG setting [12, 13], where each query is paired with a fixed context paragraph containing the target facts. The evaluation set is constructed from Wikipedia passages associated with Natural Questions [23], using named entity recognition [21] and rule-based filtering [22] to retain verifiable facts. Queries cover four fact categories: numerical facts, identifiers, temporal facts, and qualitative facts, with 250 samples per category and 1,000 paired generations in total for each experimental condition.

Models and methods. Our main experiments use LLaMA 3.1 8B Instruct [24]. For cross-model validation, we additionally evaluate Mistral-7B-Instruct-v0.3 [25] and Qwen2.5-14B-Instruct [26]. We incorporate six representative decoding-time watermarking methods: KGW [5], SWEET [17], DiPmark [18], GumbelSoft [19], SynthID-style watermarking [7], and Gumbel-Max watermarking [6]. They cover logit-bias, entropy-aware, distribution-preserving, sampling-based, and tournament-style designs. We compare the proposed FPTI, FPAI, and their joint application with two representative inference-time hallucination mitigation baselines, DoLa [27] and DeCoRe [28], under KGW watermarking. Our interventions differ in that FPTI operates on output logits and watermark perturbations, while FPAI further uses attention logits over the provided context.

Evaluation. We report the net factual accuracy decrease Δ_{fact} defined in Equation (5), with lower values indicating less watermark-induced factual degradation. Watermark detectability is measured primarily by TPR@1%FPR. Because fact-preserving interventions can alter watermark detectability, for each model, watermarking method, and intervention setting, we independently calibrate the watermark strength to restore the same measured TPR@1%FPR operating point before comparing factual accuracy. We also report perplexity and decoding speed for trade-off analysis. A supplementary package with generation, intervention, and annotation code accompanies the paper. Full experimental details are provided in Appendix B.1–B.4.

Table 1: **Effectiveness of fact-preserving interventions.** We report the net factual accuracy decrease Δ_{fact} (%) across six watermarking methods, three model families, and three matched TPR@1%FPR levels. Lower values indicate less watermark-induced factual degradation. FPTI limits factual-token perturbations, and FPTI+FPAI further reduces prefix-induced errors.

Method / setting	LLaMA3.1-8B			Mistral-7B			Qwen2.5-14B		
	TPR=.50	TPR=.90	TPR=.98	TPR=.50	TPR=.90	TPR=.98	TPR=.50	TPR=.90	TPR=.98
KGW [5]	2.1%	12.9%	22.3%	2.0%	14.4%	23.1%	2.3%	13.7%	22.7%
+FPTI	1.5%	2.4%	8.7%	1.5%	2.4%	8.9%	1.8%	2.6%	8.9%
+FPTI+FPAI	**0.6%**	**0.9%**	**6.3%**	**0.8%**	**1.2%**	**6.3%**	**0.8%**	**1.1%**	**6.4%**
SWEET [17]	2.3%	15.0%	23.1%	2.5%	15.6%	23.9%	2.9%	15.7%	24.8%
+FPTI	1.1%	1.9%	8.6%	1.3%	2.0%	8.6%	1.1%	2.1%	9.5%
+FPTI+FPAI	**0.7%**	**1.0%**	**5.5%**	**0.8%**	**1.1%**	**5.1%**	**0.7%**	**1.3%**	**6.1%**
DiPmark [18]	0.7%	6.8%	8.7%	0.7%	7.3%	9.2%	0.8%	6.5%	8.4%
+FPTI	0.4%	1.7%	4.2%	0.5%	1.9%	4.6%	0.4%	1.6%	4.3%
+FPTI+FPAI	**0.2%**	**0.4%**	**2.6%**	**0.3%**	**0.6%**	**2.7%**	**0.2%**	**0.4%**	**2.6%**
GumbelSoft [19]	2.0%	8.3%	18.2%	2.1%	8.7%	19.2%	2.7%	8.1%	19.3%
+FPTI	1.3%	3.2%	7.4%	1.3%	3.4%	7.7%	1.9%	2.8%	8.2%
+FPTI+FPAI	**0.7%**	**0.8%**	**5.6%**	**0.8%**	**1.0%**	**5.7%**	**0.7%**	**0.8%**	**5.9%**
SynthID-style [7]	0.5%	2.7%	4.8%	0.6%	2.9%	4.2%	0.6%	2.4%	4.5%
+FPTI	0.4%	0.9%	3.1%	0.4%	0.8%	3.3%	0.5%	0.8%	3.3%
+FPTI+FPAI	**0.2%**	**0.4%**	**2.1%**	**0.3%**	**0.5%**	**2.3%**	**0.3%**	**0.4%**	**2.7%**
Gumbel-Max [6]	0.6%	3.7%	5.5%	0.7%	3.6%	5.7%	0.6%	4.1%	5.2%
+FPTI	0.4%	1.1%	3.5%	0.5%	1.0%	3.7%	0.5%	1.3%	3.7%
+FPTI+FPAI	**0.2%**	**0.5%**	**2.4%**	**0.3%**	**0.6%**	**2.6%**	**0.2%**	**0.7%**	**2.5%**

5.2 Performance of Proposed Fact-Preserving Interventions

Main intervention results. We first evaluate whether the proposed interventions reduce watermarking hallucination across watermarking methods, model families, and matched detectability levels. Specifically, we compare standard watermarking with its FPTI and FPTI+FPAI variants, where FPTI removes current-step perturbations on fact-bearing tokens and FPAI further compensates for prefix-induced loss of factual attention. The results in Table 1 show a consistent reduction in Δ_{fact} from standard watermarking to FPTI and further to FPTI+FPAI. For example, under KGW on LLaMA3.1-8B at $\text{TPR@1\%FPR} = 0.90$, Δ_{fact} drops from 12.9% to 2.4% with FPTI and further to 0.9% with FPTI+FPAI. Similar reductions appear across Mistral-7B and Qwen2.5-14B, indicating that the effect is not tied to one model family. Nevertheless, residual errors remain more visible under very high detectability, suggesting that strong watermarking can push the prefix away from the evidence before later factual tokens are locally protected or attention-compensated.

The table also compares the intrinsic factuality cost of different watermarking rules under the same detectability target. The results show that watermarking methods differ substantially in how much factual degradation they introduce. For example, on LLaMA3.1-8B at $\text{TPR@1\%FPR} = 0.90$, SynthID-style watermarking has a much lower Δ_{fact} than KGW or DiPmark before intervention. This suggests that more controlled or distribution-aware sampling rules may be less prone to factual drift than stronger logit-bias-style perturbations, although all methods still benefit from FPTI+FPAI.

Comparison with general hallucination mitigation. We further compare FPTI and FPAI with DoLa [27] and DeCoRe [28] under matched KGW detectability. Table 2 reports the resulting Δ_{fact}, perplexity, and decoding-speed trade-offs, and Figure 4 visualizes the same trade-off under matched detectability. All mitigation methods improve over standard KGW by reducing both Δ_{fact} and PPL, but their trade-offs reflect different mechanisms. DoLa improves fluency: at $\text{TPR@1\%FPR} = 0.90$, it reduces PPL from 12.98 to 2.84, but Δ_{fact} remains relatively high at 9.6%. This suggests that layer-contrast decoding can make generations more fluent without directly correcting watermark-induced factual shifts. DeCoRe is more effective for factuality, lowering Δ_{fact} to 3.2%, because retrieval-head contrast is better aligned with context grounding; however, it still does not remove current-step watermark perturbations on factual tokens and incurs a larger decoding cost.

By contrast, FPTI+FPAI directly targets the two error sources identified in our mechanism analysis. FPTI removes current-step watermark perturbations on fact-bearing tokens, while FPAI compensates for prefix-induced loss of factual attention. As a result, FPTI alone reduces Δ_{fact} to 2.4%,

Table 2: **Comparing proposed interventions with hallucination mitigation baselines.** We compare our interventions with the hallucination mitigation baselines DoLa [27] and DeCoRe [28] under KGW watermarking [5]. Δ_{fact} measures the net factual accuracy decrease, and Improv. denotes its relative reduction over KGW. PPL measures fluency, and Speed is measured in generated tokens per second.

Setting	TPR@1%FPR = 0.50				TPR@1%FPR = 0.90			
	$\Delta_{\text{fact}} \downarrow$	Improv. $\uparrow$	PPL $\downarrow$	Speed $\uparrow$	$\Delta_{\text{fact}} \downarrow$	Improv. $\uparrow$	PPL $\downarrow$	Speed $\uparrow$
No watermark	0.0%	–	1.45	1.52	0.0%	–	1.45	1.52
KGW [5] (baseline)	2.1%	0.0%	3.01	1.49	12.9%	0.0%	12.98	1.48
KGW [5] + DoLa [27]	1.8%	14.3%	1.98	1.41	9.6%	25.6%	**2.84**	1.40
KGW [5] + DeCoRe [28]	1.6%	23.8%	**1.86**	1.14	3.2%	75.2%	6.74	1.10
KGW [5] + FPTI (Ours)	1.5%	28.6%	2.47	**1.48**	2.4%	81.4%	6.20	**1.47**
KGW [5] + FPTI+FPAI (Ours)	**0.6%**	**71.4%**	2.07	1.46	**0.9%**	**93.0%**	5.35	1.46

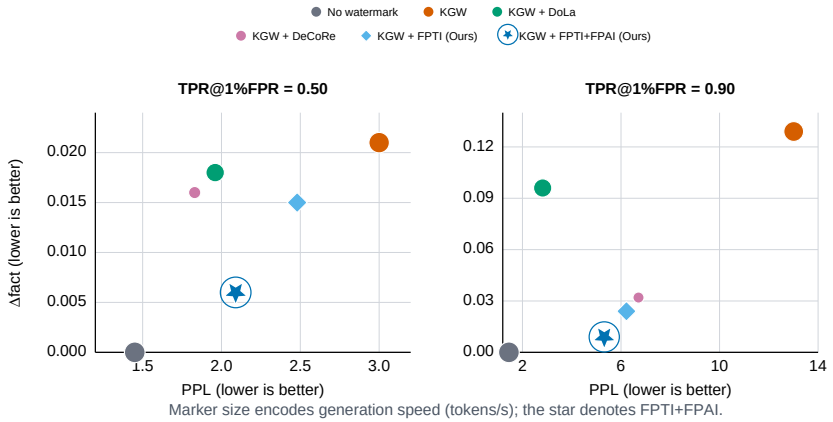


Figure 4: **Comparing trade-offs under matched detectability.** Lower-left (lower PPL and lower Δ_{fact}) is better. Marker size denotes decoding speed, with larger circles indicating higher speed.

outperforming DeCoRe in factuality with almost no speed loss; adding FPAI further reduces Δ_{fact} to 0.9%, achieving a 93.0% reduction over KGW while maintaining KGW-like speed. These results indicate that watermarking hallucination is not merely a generic hallucination or fluency problem; it is best mitigated by interventions that explicitly account for how watermarking perturbs factual token selection and evidence use.

5.3 Position-Dependent Residual Error

We use FPTI to isolate current-step token perturbations from prefix-accumulated errors. For each target query, we vary its prompt position while fixing the evidence block, target fact, template, sampling parameters, and watermark detectability. We then measure Δ_{fact} with and without FPTI, grouped by the generation position of the factual token.

Figure 5 visualizes the resulting decomposition by generation position. The gap between standard watermarking and FPTI estimates the current-step token component removed by FPTI, which remains relatively stable across positions. In contrast, the residual after FPTI grows substantially with generation depth and becomes the dominant error source after generating 70 tokens. This further supports the watermark avalanche effect: early watermark-conditioned choices gradually move the prefix away from

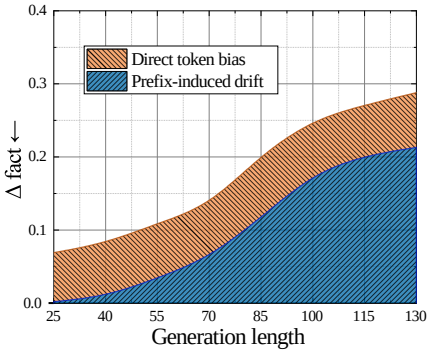


Figure 5: Position-dependent residual error under KGW. Blue: residual Δ_{fact} after FPTI; orange: reduction in Δ_{fact} by FPTI vs. standard watermarking.

the evidence-supported trajectory, so later factual decisions remain vulnerable even when their own watermark perturbations are removed.

This final analysis shows the boundary of local repair. Fact-preserving watermarking cannot rely only on protecting the current factual token; it must also limit prefix-induced drift before factual decisions are made, especially in fact-critical applications where small substitutions in dates, identifiers, or numerical values can be consequential.

6 Discussion and Conclusion

This paper studies factual degradation caused by watermarking itself. We identify *watermarking hallucination* as a failure mode distinct from ordinary hallucination and surface-quality loss: watermarked text can remain fluent and detectable while changing a context-supported fact. Using the net factual accuracy decrease Δ_{fact}, we show that this effect appears across watermarking methods, fact types, and model families. We attribute it to two coupled sources: current-step token perturbations and prefix-induced attention interference. FPTI and FPAI validate this mechanism by protecting fact-bearing output tokens and restoring attention to factual context, substantially reducing watermark-induced errors.

Our study also reveals the boundary of local repair. Under strong watermarking, residual errors remain. Early watermark-conditioned choices can push the prefix away from the supporting evidence before later factual decisions occur. This suggests that factuality should be treated as a watermark design constraint, not only a post-hoc metric. Future watermarks should place their signal where it least interferes with grounding, avoiding numerical values, identifiers, dates, citations, and prefix positions that affect later factual attention. Overall, watermarking should not be assumed harmless in fact-critical applications: factual reliability must be evaluated alongside detectability, robustness, and output quality.

Acknowledgments

We thank the anonymous reviewers for their helpful suggestions. This work was supported by the National Key Research and Development Program of China under Grant No. 2023YFB2703900. The authors declare no competing interests.

References

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

[2] Yisroel Mirsky, Ambra Demontis, Jaidip Kotak, Ram Shankar, Deng Gelei, Liu Yang, Xiangyu Zhang, Maura Pintor, Wenke Lee, Yuval Elovici, and Battista Biggio. The threat of offensive ai to organizations. *Computers & Security*, 124:103006, 2023. ISSN 0167-4048. doi: 10.1016/j.cose.2022.103006. URL https://www.sciencedirect.com/science/article/pii/S0167404822003984.

[3] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93, 01 2024. ISSN 2161-7201. doi: 10.1093/jla/laae003. URL https://doi.org/10.1093/jla/laae003.

[4] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242, 2025. doi: 10.1111/jels.12413. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jels.12413.

[5] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In Andreas Krause, Emma Brunskill,

Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html.

[6] Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=FpaCL1MO2C.

[7] Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, Jamie Hayes, Nidhi Vyas, Majd Al Merey, Jonah Brown-Cohen, Rudy Bunel, Borja Balle, Taylan Cemgil, Zahra Ahmed, Kitty Stacpoole, Ilia Shumailov, Cip Baetu, Sven Gowal, Demis Hassabis, and Pushmeet Kohli. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024. doi: 10.1038/s41586-024-08025-4. URL https://doi.org/10.1038/s41586-024-08025-4.

[8] Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. A survey of text watermarking in the era of large language models. *ACM Comput. Surv.*, 57(2), November 2024. ISSN 0360-0300. doi: 10.1145/3691626. URL https://doi.org/10.1145/3691626.

[9] Zhiguang Yang, Gejian Zhao, and Hanzhou Wu. Watermarking for large language models: A survey. *Mathematics*, 13(9), 2025. ISSN 2227-7390. doi: 10.3390/math13091420. URL https://www.mdpi.com/2227-7390/13/9/1420.

[10] Eva Giboulot and Teddy Furon. Watermax: breaking the llm watermark detectability-robustness-quality trade-off. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 18848–18881. Curran Associates, Inc., 2024. doi: 10.52202/079017-0597. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/21b5883bc8fec922fdbbb06675388164-Paper-Conference.pdf.

[11] Rochana Prih Hastuti, Rian Adam Rajagede, Mansour Al Ghanim, Mengxin Zheng, and Qian Lou. Factuality beyond coherence: Evaluating LLM watermarking methods for medical texts. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15129–15147, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.818. URL https://aclanthology.org/2025.findings-emnlp.818/.

[12] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. URL https://arxiv.org/abs/2312.10997.

[13] Bo Ni, Zheyuan Liu, Leyao Wang, Yongjia Lei, Yuying Zhao, Xueqi Cheng, Qingkai Zeng, Luna Dong, Yinglong Xia, Krishnaram Kenthapadi, Ryan Rossi, Franck Dernoncourt, Md Mehrab Tanjim, Nesreen Ahmed, Xiaorui Liu, Wenqi Fan, Erik Blasch, Yu Wang, Meng Jiang, and Tyler Derr. Towards trustworthy retrieval augmented generation for large language models: A survey, 2025. URL https://arxiv.org/abs/2502.06872.

[14] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2), January 2025. ISSN 1046-8188. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155.

[15] Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

[16] Yufei Tao, Adam Hiatt, Erik Haake, Antonie J. Jetter, and Ameeta Agrawal. When context leads but parametric memory follows in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4034–4058, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.234. URL https://aclanthology.org/2024.emnlp-main.234/.

[17] Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoo Yun, Jamin Shin, and Gunhee Kim. Who wrote this code? watermarking for code generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4890–4911, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.268. URL https://aclanthology.org/2024.acl-long.268/.

[18] Yihan Wu, Zhengmian Hu, Junfeng Guo, Hongyang Zhang, and Heng Huang. A resilient and accessible distribution-preserving watermark for large language models, 2024. URL https://arxiv.org/abs/2310.07710.

[19] Jiayi Fu, Xuandong Zhao, Ruihan Yang, Yuansen Zhang, Jiangjie Chen, and Yanghua Xiao. GumbelSoft: Diversified language model watermarking via the GumbelMax-trick. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5791–5808, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.315. URL https://aclanthology.org/2024.acl-long.315/.

[20] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[21] Zhuojun Ding, Wei Wei, and Chenghao Fan. Selecting and merging: Towards adaptable and scalable named entity recognition with large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9869–9886, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.487. URL https://aclanthology.org/2025.acl-long.487/.

[22] Angel X. Chang and Christopher D. Manning. TokensRegex: Defining cascaded regular expressions over tokens. Technical Report CSTR 2014-02, Department of Computer Science, Stanford University, 2014.

[23] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

[24] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[25] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

[26] Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

[27] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Th6NyL07na.

[28] Aryo Pradipta Gema, Chen Jin, Ahmed Abdulaal, Tom Diethe, Philip Alexander Teare, Beatrice Alex, Pasquale Minervini, and Amrutha Saseendran. DeCoRe: Decoding by contrasting retrieval heads to mitigate hallucinations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10003–10039, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.531. URL https://aclanthology.org/2025.findings-emnlp.531/.

[29] Xinwei Zhang, Haibo Hu, Qingqing Ye, Li Bai, and Huadi Zheng. MER-Inspector: Assessing model extraction risks from an attack-agnostic perspective. In *Proceedings of the ACM Web Conference 2025*, pages 4300–4315, 2025. doi: 10.1145/3696410.3714894.

[30] Aoting Hu, Zhigang Lu, Renjie Xie, and Minhui Xue. VeriDIP: Verifying ownership of deep neural networks through privacy leakage fingerprints. *IEEE Transactions on Dependable and Secure Computing*, 21(4):2568–2584, 2024. doi: 10.1109/TDSC.2023.3313577.

[31] Xinwei Zhang, Hangcheng Liu, Li Bai, Hao Wang, Qingqing Ye, Tianwei Zhang, and Haibo Hu. On the adversarial robustness of large vision-language models under visual token compression. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=q5sBPXOuHC.

[32] Liming Lu, Shuchao Pang, Jiakai Wang, Xiang Gu, Yunhuai Liu, Xianglong Liu, and Yongbin Zhou. QuEST: Quantization-conditioned efficient stealthy trojan. *IEEE Transactions on Information Forensics and Security*, 21:2962–2977, 2026. doi: 10.1109/TIFS.2026.3671079.

[33] Shuchao Pang, Xiyu Zeng, Siyuan Liang, Liming Lu, Haotian Zhu, Chuanting Zhang, Enguang Liu, Liang Zhang, Basem Shihada, Yongbin Zhou, and Jason Xue. SafeSteer: Adaptive subspace steering for efficient jailbreak defense in vision language models. *IEEE Transactions on Information Forensics and Security*, 21:7157–7172, 2026. doi: 10.1109/TIFS.2026.3714149.

[34] Mingjia Huo, Sai Ashish Somayajula, Youwei Liang, Ruisi Zhang, Farinaz Koushanfar, and Pengtao Xie. Token-specific watermarking with enhanced detectability and semantic coherence for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[35] Haiyun He, Yepeng Liu, Ziqiao Wang, Yongyi Mao, and Yuheng Bu. Theoretically grounded framework for LLM watermarking: A distribution-adaptive approach. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=CMmKcHFDKL.

[36] Saksham Rastogi and Danish Pruthi. Revisiting the robustness of watermarking to paraphrasing attacks. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18100–18110, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1005. URL https://aclanthology.org/2024.emnlp-main.1005/.

[37] Qilong Wu and Varun Chandrasekaran. Bypassing LLM watermarks with color-aware substitutions. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8549–8581, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.464. URL https://aclanthology.org/2024.acl-long.464/.

[38] Nikola Jovanović, Robin Staab, and Martin Vechev. Watermark stealing in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[39] Abdulrahman Diaa, Toluwani Aremu, and Nils Lukas. Optimizing adaptive attacks against watermarks for language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 13546–13569. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/diaa25a.html.

[40] Guangzhi Xiong, Zhenghao He, Bohan Liu, Sanchit Sinha, and Aidong Zhang. Toward faithful retrieval-augmented generation with sparse autoencoders. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hgBZP67BkP.

[41] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

[42] Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Zixia Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V Le. Long-form factuality in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=4M9f8VMt2C.

[43] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

[44] Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. Knowledge boundary of large language models: A survey. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5131–5157, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.256. URL https://aclanthology.org/2025.acl-long.256/.

[45] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 41451–41530. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/81b8390039b7302c909cb769f8b6cd93-Paper-Conference.pdf.

[46] Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=uWVC5FVidc.

[47] Tom Sander, Hongyan Chang, Tomáš Souček, Tuan Tran, Valeriu Lacatusu, Sylvestre-Alvise Rebuffi, Alexandre Mourachko, Surya Parimi, Christophe Ropers, Rashel Moritz, Vanessa Stark, Hady Elsahar, and Pierre Fernandez. Textseal: A localized llm watermark for provenance & distillation protection, 2026. URL https://arxiv.org/abs/2605.12456.

[48] Zongqi Wang, Tianle Gu, Baoyuan Wu, and Yujiu Yang. Morphmark: Flexible adaptive watermarking for large language models, 2025. URL https://arxiv.org/abs/2505.11541.

A Related Works

The growing deployment of AI systems has brought increasing attention to their security, including model ownership protection and robustness against adversarial threats [29–33]. As AI-generated content becomes more widespread, tracing its provenance is another important concern. Text watermarking offers a means of identifying model-generated content, but its effects on factual reliability also require careful evaluation. We review the literature on LLM watermarking, factuality risks in watermarked text, context-grounded generation and evaluation, and decoding-time methods for improving factuality.

LLM watermarking. LLM watermarking methods embed detectable signals into generated text by modifying decoding, constraining sampling, or preserving distributional properties while adding a hidden statistical structure [5–9]. KGW-style methods bias token sets directly, while methods such as DiPmark, SWEET, GumbelSoft, token-specific watermarking, and distribution-adaptive watermarking seek better quality and robustness trade-offs [18, 17, 19, 34, 35]. Robustness studies examine paraphrase attacks, watermark removal, watermark stealing, and adaptive evasion attacks [36–39]. These approaches are usually evaluated through detectability, false positive rate, robustness to edits, and surface quality. Our work studies a different axis: whether the same decoding changes also harm factual reliability.

Factuality risks in watermarked text generation. Hastuti et al. [11] show that coherence-oriented evaluation is insufficient for medical watermarked generation and introduce a GPT-based judging workflow for factuality-aware assessment. Their results suggest that factual degradation can be more severe than coherence loss, especially when watermarking shifts entity-level or domain-specific content. Our work builds on this observation but differs in scope and goal. First, we study watermarking hallucination under a paired context-grounded setting, where the relevant evidence is explicitly available and the unwatermarked model can answer correctly. Second, rather than measuring factuality as an absolute output property, we use Δ_{fact} to measure the net factual accuracy decrease between paired unwatermarked and watermarked generations. Third, we go beyond evaluation by attributing the degradation to current-step token perturbations and prefix-induced attention interference, and by introducing fact-preserving interventions that target these mechanisms.

Context-grounded factuality and RAG evaluation. Retrieval-augmented generation is widely used to reduce hallucination by conditioning generation on external evidence [12, 13]. However, the presence of relevant context does not by itself guarantee faithful output: models may ignore the supplied evidence, rely on parametric priors, misread retrieved passages, or drift away from the context during long-form generation [16, 15]. Recent work further studies faithful RAG through internal model signals, for example using sparse autoencoders to identify features associated with source-faithful and unfaithful generation [40]. Factuality evaluation methods therefore increasingly emphasize evidence-grounded verification rather than surface fluency alone. FActScore decomposes long-form answers into atomic facts and checks whether they are supported [41], while SAFE uses structured search and verification to evaluate long-form factuality [42]. Our evaluation follows the same evidence-grounded principle but applies it to a different causal question: given the same model, prompt, retrieved context, and sampling configuration, does watermarking itself introduce factual errors? This paired design separates watermark-induced factual degradation from ordinary model errors or retrieval failures.

Hallucination and decoding-time factuality. Hallucination research identifies failures caused by missing knowledge, knowledge conflict, weak grounding, uncertainty, and inference-time dynamics [14–16, 43]. A complementary line of work studies the knowledge boundary of LLMs, where factual failures arise because models cannot reliably memorize, access, or apply certain types of knowledge [44]. Decoding interventions can improve or degrade factuality, as shown by methods such as ITI and DoLa [45, 27]. These methods treat hallucination as a general generation failure and aim to make the model more faithful. In contrast, *watermarking hallucination* is induced by a source-attribution mechanism inserted for detectability: watermark-conditioned token choices can alter the generated prefix and later weaken attention to factual evidence. This makes *watermarking hallucination* a distinct inference-time failure mode: the model may be capable of using the evidence under clean decoding, but the watermark shifts the generation trajectory away from that evidence. Our FPTI and FPAI interventions are therefore not general hallucination mitigators; they are watermark-aware probes and repairs designed to separate and reduce the two mechanisms responsible for watermark-induced factual errors.

B Experimental Details

B.1 Model and Decoding Setup

We use instruction-tuned open-weight models throughout the experiments. LLaMA 3.1 8B Instruct [24] is used as the main model for the mechanism analysis and the primary watermarking evaluation. To test whether the observed trends generalize across model families, we additionally evaluate Mistral-7B-Instruct-v0.3 [25] and Qwen2.5-14B-Instruct [26] in the cross-model intervention study.

For watermarking, we evaluate six representative decoding-time watermarking paradigms: KGW [5], SWEET [17], DiPmark [18], GumbelSoft [19], SynthID-style watermarking [7], and Gumbel-Max-style watermarking [6]. Our implementations follow their corresponding published mechanisms and cover logit-bias, entropy-aware, distribution-preserving, sampling-path, and tournament-style designs. For mitigation, we evaluate the proposed FPTI and FPAI interventions. In the comparison with general hallucination mitigation methods, we further include DoLa [27] and DeCoRe [28] under KGW watermarking.

For each condition, the unwatermarked and watermarked outputs are generated with the same prompt, context, sampling parameters, and random-seed schedule. This pairing is important because the goal is not to estimate general model factuality, but to isolate the factual change caused by adding a watermark. When comparing watermarking methods or interventions, we align settings by watermark detectability whenever possible rather than by nominal strength, since different watermarking rules can produce different detector performance under the same strength parameter.

Licenses and existing assets. We use publicly available datasets, open-weight models, and open-source software packages. The RAG contexts are constructed from Wikipedia passages associated with Natural Questions [23], whose official repository is released under Apache-2.0. LLaMA 3.1 8B Instruct is used under the Meta Llama 3.1 Community License. Mistral-7B-Instruct-v0.3 and Qwen2.5-14B-Instruct are released under Apache-2.0. For watermarking and decoding baselines, we cite the corresponding papers and use either publicly available implementations when their licenses permit research use or our own implementations based on the published descriptions. We do not redistribute model weights or original dataset contents. The supplementary package provides paired-generation scripts, intervention-generation scripts, structured annotation utilities, and toy examples for smoke testing; full experiments require evaluation samples following the JSONL schema described in the package.

Table 3 summarizes the decoding configuration used for unwatermarked, watermarked, and intervention-enhanced generations.

Table 3: **Decoding configuration.** The same decoding parameters are used for unwatermarked, watermarked, and intervention-enhanced generations unless an experiment explicitly varies watermark strength or detectability.

Parameter	Value	Parameter	Value
Temperature	0.7	Top-p	0.9
Top-k	40	Repetition penalty	1.1
No-repeat n-gram size	3	Maximum input context	300 tokens
Maximum new tokens	300	Sampling	Enabled

B.2 RAG Evaluation Setup

We construct a controlled retrieval-augmented generation (RAG) setting, where each query is paired with a fixed context paragraph containing multiple verifiable facts.

The evaluation set is built from Wikipedia passages associated with Natural Questions. We apply named entity recognition (NER) together with rule-based filtering to identify candidate contexts, and retain those that contain at least three facts of the same type. The resulting contexts span diverse topics (e.g., legal, medical, and financial domains) and cover four categories of factual information.

Each evaluation sample is designed to target a specific category of facts. For each query, the model is required to extract and answer one or more target facts of the same type from the provided context within a single response.

We evaluate four categories of factual information, covering different types of reasoning and representation:

Numerical facts refer to quantitative values (e.g., "78 mg/dL"), which require precise numerical correctness and sensitivity to units.

Identifier facts include named entities, codes, or symbolic references (e.g., "AAPL", "Doe v. Smith"), where correctness depends on exact matching.

Temporal facts correspond to dates and time expressions (e.g., "May 12"), which must be chronologically accurate.

Qualitative facts represent descriptive or evaluative information (e.g., "relatively minor"), which involves comparative or contextual interpretation.

For each fact type, we construct 250 evaluation samples, resulting in 1,000 paired generations for each experimental condition and a balanced test set across factual categories.

Data source and privacy. Our evaluation data are constructed from publicly available Wikipedia passages associated with Natural Questions. We do not collect new user data or private information. The dataset is used only for controlled factuality evaluation, and each item is derived from a public context paragraph paired with a synthetic or filtered query. Although the contexts cover diverse topics, the benchmark is not intended to represent all fact-critical domains, such as clinical, legal, or financial deployment settings.

B.3 Net Factual Accuracy Decrease Evaluation Protocol

We evaluate watermark-induced factual degradation using the paired metric defined in Equation (5). For each evaluation item, we compare the unwatermarked output with the corresponding watermarked output under the same context, query, sampling configuration, and random-seed schedule. We record both correct-to-incorrect and incorrect-to-correct transitions and compute Δ_{fact} as their difference over the full evaluation set.

The factuality evaluator checks whether the answer is supported by the provided context. For numerical, temporal, and identifier facts, we use exact or normalized matching whenever possible, allowing harmless surface variations such as equivalent date formats or unit spellings. For qualitative facts, we use structured annotation followed by manual verification on sampled outputs. A watermarked answer is marked incorrect if it omits a required target fact, changes a context-supported value, or introduces a contradictory claim.

This paired protocol differs from absolute factuality judging. It does not ask whether a model is generally factual across open-ended tasks. Instead, it measures the change in factual accuracy caused by watermarking while accounting for transitions in both directions.

We do not rely on an LLM judge for the main reported factuality labels. Rule-based decisions are manually audited on sampled outputs to check whether normalization and matching rules correctly capture factual support.

B.4 Attention Measurement and Detectability

To compute the factual attention coefficient α_t, we first identify fact-bearing context tokens associated with the target facts. We then aggregate attention mass assigned to those tokens across layers and heads at each decoding step. The resulting α_t is averaged over generated samples within the same watermarking method and detectability regime. This measurement is not used as a standalone explanation of model reasoning; it is used as an observable signal whose causal relevance is tested through FPAI in the main text.

Detectability is aligned by grouping runs into low-, mid-, and high-Z regimes when the detector reports a Z-score. This avoids comparing methods at the same nominal hyperparameter when the

actual watermark signal differs. For methods without a directly comparable Z-score, appendix plots use the method's native strength axis.

B.5 Representative Intervention Cases

Table 4 shows a case where the combined intervention restores the target facts. The standard watermark produces plausible but wrong numerical values. FPTI+FPAI recovers the original values because the early part of the answer still follows the evidence closely enough for attention restoration to work.

Table 4: **Successful recovery under FPTI+FPAI.** The standard watermark produces plausible but incorrect numerical values, while FPTI+FPAI restores the context-supported answer.

Setting	Output	Factual?
No watermark	Fasting glucose range: 145–162 mg/dL; postprandial peak: 210 mg/dL; nocturnal low: 78 mg/dL.	Yes
Watermark	Fasting glucose ranged 145–160 mg/dL; peak value 200–210 mg/dL; nocturnal low around 73 mg/dL.	No
FPTI+FPAI	Fasting sugar level: 145–162 mg/dL; peak value: 210 mg/dL; nocturnal low: 78 mg/dL.	Yes

Table 5 shows the limit of local repair. Once the watermarked prefix becomes malformed, restoring factual tokens and attention can recover part of the answer but cannot reconstruct the full unwatermarked trajectory.

Table 5: **Incomplete recovery under FPTI+FPAI.** Severe prefix drift leaves residual factual errors after both interventions.

Setting	Output	Factual?
No watermark	Fasting glucose range: 145–162 mg/dL; postprandial peak: 210 mg/dL; nocturnal low: 78 mg/dL.	Yes
Watermark	Fasting glucrose range for this particular patient span from approximately between 145–155 to 148 milligrm/dltrange.	No
FPTI+FPAI	Fasting sugar readings' range: 145–162 mg/dL. Specific value was not listed.	No

B.6 Compute Resources

All experiments are performed on a cluster equipped with NVIDIA GeForce RTX 4090 GPU. A single strength-setting run took approximately 70,000–110,000 seconds. For each watermarking method, the cross-model mitigation sweep over 8 strengths, 3 settings, and 3 models required approximately 560–880 GPU hours. The total compute varies depending on the number of watermarking methods included in each reported experiment; preliminary debugging and failed runs are not included in this estimate.

B.7 Code and Data Availability

We provide a supplementary package containing code for paired RAG generation, watermark detectability logging, FPTI/FPAI intervention generation, and structured annotation aggregation. The code is available at `https://github.com/Haocheng-NJUST/WM-Hallucination`. The package reads evaluation items from a JSONL schema containing the context, query, fact type, and target facts. Generation scripts do not infer final factuality labels from raw text; final labels are loaded from structured annotation files, with qualitative cases manually verified as described above. The package includes toy examples for smoke testing the pipeline. We do not redistribute model weights or original dataset contents; users should obtain open-weight models and public datasets from their official sources.

Table 6: **Factual transitions for additional watermark families.** Correct-to-incorrect and incorrect-to-correct transitions use the full set of 1,000 paired generations as their denominator. The net factual accuracy decrease is their difference.

Method	Correct→Incorrect	Incorrect→Correct	Δ_{fact} [95% CI]
Unbiased Watermark (δ-reweighting)	7.1%	3.5%	3.6% [3.1%, 3.9%]
TextSeal	7.7%	3.8%	3.9% [1.8%, 5.8%]
MorphMark	16.4%	4.3%	12.1% [11.5%, 13.7%]

Table 7: **Fluent does not imply factual.** The watermarked output by SynthID [7] has *lower perplexity* but introduces an entity-level factual error, while the non-watermarked output remains correct. In our evaluation under the standard SynthID-style setting, such entity-level watermarking hallucinations occur in 1.7% of cases.

Question: What case sets the key precedent for liability standards in the Ninth Circuit?

Context (RAG)	**No Watermark**	**With Watermark (SynthID)**
... **Legal context** *Precedent:* **Doe v. Smith** *Citation:* 2023 WL 183472 (9th Cir.) *Summary:* Establishes supervisory liability standards in the Ninth Circuit. ...	The controlling precedent is **Doe v. Smith**, 2023 WL 183472 (9th Cir.), which establishes the framework for supervisory liability. *Factually correct* **PPL = 4.17**	The controlling authority is **Doe v. Smithson**, 2023 WL 183472 (9th Cir.). *Entity-level factual error* **PPL = 3.31 (lower)**

C Additional Analysis

C.1 Additional Watermark Families

We further evaluate Unbiased Watermark with δ-reweighting [46], TextSeal [47], and MorphMark [48] to cover additional watermark designs. For each method, we evaluate 1,000 paired generations covering four fact categories, with 250 examples per category. The watermark strength is calibrated to a measured $\text{TPR@1\%FPR} = 0.90$.

All three methods produce transitions in both directions, although the incorrect-to-correct transitions do not fully offset the correct-to-incorrect transitions. Unbiased Watermark and TextSeal yield relatively small net factual accuracy decreases of 3.6 and 3.9 percentage points, respectively, whereas MorphMark produces a substantially larger decrease of 12.1 points. Together with the results in Figure 1 and Table 1, these findings show that the magnitude of watermark-induced factual degradation depends on the watermarking design.

C.2 Fluency Metrics and Factual Errors

We examine whether token-level perplexity separates factually correct and incorrect tokens. For each generated token with a factual correctness label, we compute its perplexity under the base model. Figure 6 shows that incorrect factual tokens are not restricted to high-perplexity regions. Correct and incorrect tokens overlap substantially, including in the low-perplexity region highlighted by the inset.

This analysis explains the example in Table 7. A wrong token can be easy for the language model to generate if it fits the local sentence pattern. Factuality depends on whether the token matches the supplied evidence, not only on whether the token is likely under the model. Watermark evaluation therefore needs paired factuality measures such as Δ_{fact} in addition to detectability and fluency checks.

C.3 Entropy Does Not Identify Factual Tokens

SWEET avoids watermarking low-entropy positions, which makes it a natural baseline for reducing quality degradation. However, low entropy is not the same as factual importance. A token can be factual even when the model assigns high uncertainty to it, especially for numerical values, identifiers, and dates where multiple surface forms may be locally plausible but only one is supported by the context.

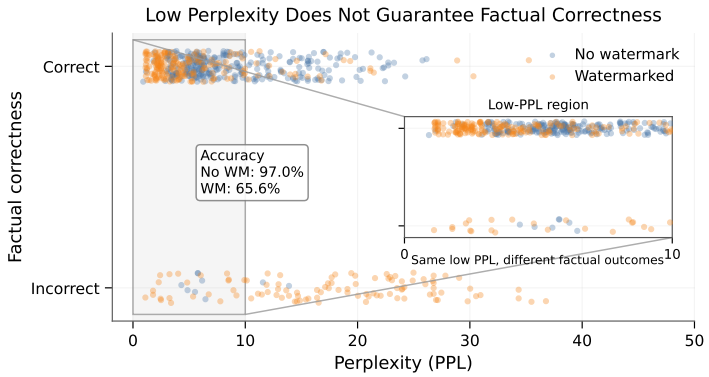


Figure 6: **Token-level perplexity versus factual correctness.** Many incorrect factual tokens appear in low-perplexity regions, showing why fluency-oriented metrics can miss watermark-induced factual errors.

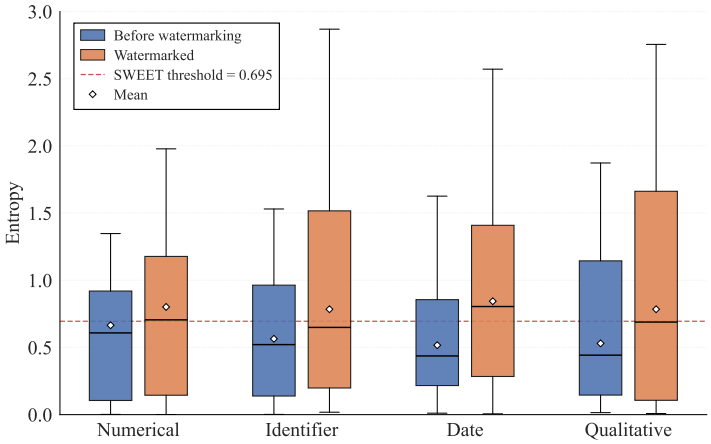


Figure 7: **Entropy of factual tokens before and after watermarking.** We report the mean token-level entropy for factual tokens across four fact types. The dashed line marks the SWEET entropy threshold (0.695). Many factual tokens lie above this threshold, indicating that entropy-based protection does not reliably cover all fact-bearing tokens.

As shown in Figure 7, factual tokens are not confined to the low-entropy region protected by SWEET. This explains why entropy-aware watermarking can still produce watermarking hallucination: protecting low-entropy positions preserves some high-confidence generation choices, but many factual decisions occur at higher-entropy positions where the watermark remains active. Consequently, SWEET reduces perturbations for a subset of tokens that are important for generation quality, but it does not provide explicit protection for all fact-bearing tokens. This suggests that entropy-aware protection can reduce some generation-quality perturbations, but it does not by itself guarantee fact preservation.

D Broader Impact

This work can contribute to a safer and more responsible deployment of LLM watermarking by revealing a factual reliability risk that is not captured by standard detectability, robustness, or fluency metrics. Our study does not argue against watermarking. Instead, it shows that watermark deployment in fact-critical applications should be accompanied by factuality evaluation and domain-specific validation. By identifying watermark-induced hallucination and proposing fact-preserving interventions, our study can help model providers, policymakers, and application developers evaluate watermarking systems more comprehensively before deploying them in fact-critical settings. The

proposed analysis may also encourage future watermark designs to preserve factual reliability while maintaining provenance and accountability benefits. This work may also have unintended negative impacts. The finding that watermarking can induce factual errors could be misinterpreted as evidence that watermarking is broadly unreliable, potentially weakening trust in legitimate provenance mechanisms. In addition, our failure-mode analysis may help adversarial users identify settings where watermarked generation is more vulnerable to factual errors, such as strong watermarking or long generations. These risks highlight the need to interpret our findings as a call for more careful watermark evaluation and deployment, rather than as a rejection of watermarking itself.

E Limitations

Although our interventions reduce watermark-induced hallucinations, they do not completely eliminate factual degradation under strong watermarking or long generations, because a realized watermark-conditioned trajectory can still diverge from the evidence-supported trajectory. In addition, our evaluation is conducted in a controlled RAG setting with selected fact types, and may not cover all forms of factuality errors in open-ended long-form generation. Future work should study more diverse domains, longer outputs, human-validated factuality evaluation, and watermark objectives that explicitly incorporate factuality-preserving constraints.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's main scope: studying watermark-induced factual degradation, analyzing token-level and attention-level mechanisms, and evaluating fact-preserving interventions under controlled RAG settings. The claims are supported by the empirical results and mechanism analyses in the main paper, with limitations and societal implications discussed in the appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in Appendix E, including residual risks under strong watermarking and long generations, the fact that the interventions reduce but do not eliminate factual degradation, and the need to evaluate watermarking beyond detectability and robustness in fact-critical settings.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present formal theoretical theorems or proofs; its contributions are methodological and empiric.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the evaluated models, decoding parameters, watermarking methods, paired net factual accuracy decrease evaluation protocol, RAG construction procedure, factual categories, attention measurement, and detectability alignment details in Appendix B. We also provide code and reproduction scripts at `https://github.com/Haocheng-NJUST/WM-Hallucination` to support reproduction of the main experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide a public code repository at `https://github.com/Haocheng-NJUST/WM-Hallucination`, including paired-generation scripts, intervention-generation scripts, structured annotation utilities, dependency specifications, and toy examples for smoke testing. The repository documents the JSONL schema required to run the full evaluation and explains that model weights and original dataset contents are not redistributed.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the evaluated models, watermarking methods, decoding hyperparameters, watermark strength sweeps, paired net factual accuracy decrease evaluation protocol, RAG evaluation construction, factual categories, number of samples, attention measurement, and detectability alignment in Appendix B. Additional runnable configurations are provided in the public code repository.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All relative watermarking hallucination rate (net factual accuracy decrease) results are reported under fixed watermark detectability operating points, primarily TPR@1%FPR, so comparisons are made at matched detectability rather than nominal watermark strength. In Figure 1(a), error bars denote bootstrap confidence intervals computed over paired evaluation samples under each watermarking configuration, quantifying variability in the main factual-degradation trends.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that would be out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We report the GPU type, memory, per-run runtime, and estimated GPU hours for the main cross-model mitigation sweep in Appendix B. All experiments are performed on a cluster equipped with NVIDIA GeForce RTX 4090 GPU, and a full sweep for each watermarking method required approximately 560–880 GPU hours.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conforms to the NeurIPS Code of Ethics. The study uses publicly available datasets and open-weight models for controlled factuality evaluation, does not involve human subjects or private user data.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss broader impacts in Appendix D. The work can help researchers and practitioners identify factual reliability risks in watermark deployment and motivate factuality-aware evaluation, while also noting that emphasizing such failures may affect confidence in watermarking systems if the results are interpreted without their controlled experimental scope.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a new pretrained model, image generator, or high-risk scraped dataset. The public repository contains code and reproduction scripts for controlled evaluation, and we do not redistribute model weights or original dataset contents.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original papers and sources for the datasets, models, watermarking methods, and software libraries used in the experiments. We document the relevant licenses or terms of use in the appendix and repository README, and we do not redistribute model weights or original dataset contents.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide code, reproduction scripts, configuration files, and documentation in the public repository. The paper and repository README describe the evaluation protocol, data construction procedure, intended use, and limitations of the released materials.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects. Factuality labels are obtained through rule-based matching and author-side manual audits rather than participant studies or paid annotation.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects, so IRB approval or equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are the evaluated systems in this paper, but they are not used as an important, original, or non-standard component of the core research methodology. The main net factual accuracy decrease labels are obtained through rule-based matching and author-side manual audits rather than LLM-based judging.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.